Fetal Health Classification

Présentation de l'étude

Contexte:

Chaque année:

- 290 000 décès maternels
- 1.9 million de morti-naissances
- 2,3 millions de décès de nouveau-nés

Le **cardiotocogramme**, un examen qui enregistre la fréquence cardiaque fœtale et l'activité utérine de la mère. Une meilleure interprétation → une meilleure prise en charge.

Jeu de données :

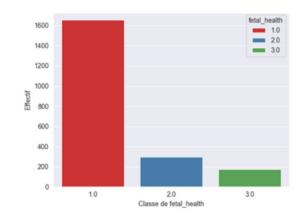
22 variables qui décrivent **2126 cardiotocogrammes**. Ces variables décrivent :

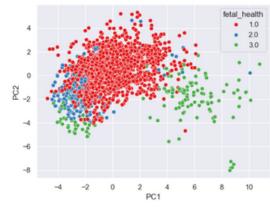
- l'histogramme des rythmes cardiaques du fœtus
- l'activité utérine de la mère (contractions, mouvements du fœtus...)
- Chaque individu a été classé parmi trois catégories : **Normal, Suspect, Pathologique**

Objectif: Prédire à quelle classe un individu appartient, tout en limitant les erreurs.

Exploration des données

- Variables prédictives : 21 variables quantitatives
- Variable à prédire : fetal health qui est qualitative ordinale:
 1 (Normal) < 2 (Suspect) < 3 (Pathological)





 $\begin{array}{ll} {\rm FIGURE} \ 1 - {\rm Proportion} \ {\rm des} \ {\rm classes} \ {\rm de} \ {\rm la} \ {\rm variable} \ {\rm \grave{a}} \ {\rm pr\acute{e}} \\ {\rm dire}. \end{array} \qquad \begin{array}{ll} {\rm FIGURE} \ 2 - {\rm Repr\acute{e}sentation} \ {\rm des} \ {\rm donn\acute{e}es} \ {\rm dans} \ {\rm le} \ {\rm premier} \\ {\rm plan} \ {\rm factoriel} \ ({\rm ACP} \ {\rm normalis\acute{e}e} \ {\rm et} \ {\rm pond\acute{e}r\acute{e}e}) \end{array}$

Classification supervisée

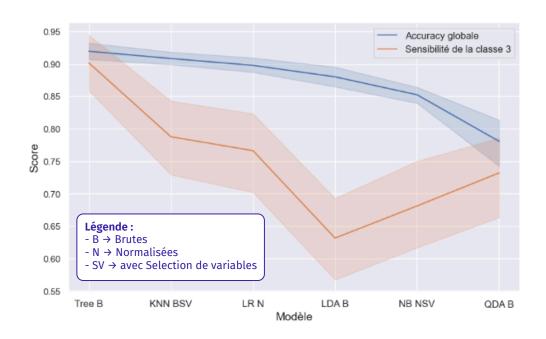


FIGURE 3 – Performances des différents classifieurs

Prétraitements

Sélection des variables

Afin d'éviter le fléau de la dimension, une sélection des variables est faite en réalisant une **ANOVA** entre les variables et fetal health.

Un F Score élevé indique une forte relation entre la variable et la variable cible.

Table 1 – Variables les plus liées à l'état de santé du fœtus (F score > 200).

Variable	F Score	
prolongued decelerations	505	
abnormal_short_term_variability	344	
% time abnormal long term variability	345	
histogram_mode	275	
histogram mean	298	
histogram_median	249	

Normalisation

Afin, que le poids d'une variable dans la création du modèle ne soit pas influencé par sa variance, il est important de normaliser les données. Cela permet d'éviter les biais d'échelles.

Nos meilleurs classifieurs

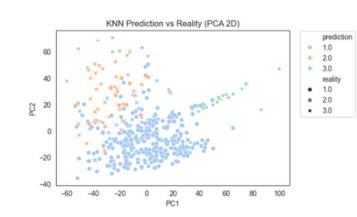


FIGURE 4 – Prédiction KNN VS Realité (représentation dans le premier plan factoriel)

K Plus Proches Voisins:

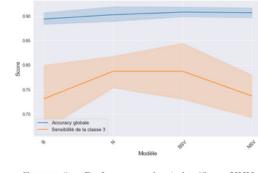
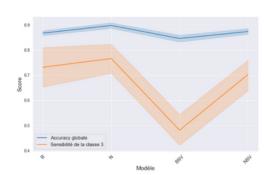


Figure 5 – Performances des 4 classifieurs KNI

Table 2 – Performances du classifieur KNN avec séléction de variables (moyennes des 10 plis de notre validation croisée)

Classe	Précision	Rappel	F1-score
Normal	0.94	0.96	0.95
Suspect	0.68	0.68	0.67
Pathological	0.91	0.79	0.84
Global	0.91		

Régression logistique :



 $\begin{tabular}{ll} Figure 6-Performances des 4 modèles de Régression \\ Logistique \\ \end{tabular}$

Table 3 – Performances de la régression logistique avec normalisation des données (moyennes des 10 plis de notre validation croisée)

Classe	Précision	Rappel	F1-score
Normal	0.95	0.95	0.95
Suspect	0.67	0.67	0.66
Pathological	0.85	0.77	0.80
Global	0.90		

Arbres de décision :

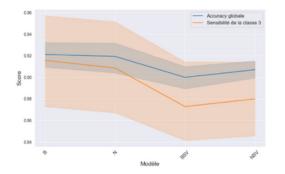


FIGURE 7 – Performances des 4 modèles d'Arbre de Décision

Table 4 – Performances de l'arbre de décision avec les données brutes (moyennes des 10 plis de notre validation croisée)

Classe	Précision	Rappel	F1-score
Normal	0.96	0.95	0.96
Suspect	0.75	0.76	0.76
Pathological	0.91	0.91	0.91
Global	0.92		